混合高斯分布

518030910150 方泓杰

Nov. 2nd, 2019

1 问题简述

定义 1 (混合高斯分布) 设$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$，则将$Z = X + \eta Y$ 服从的分布称为参数为$\mu_1, \sigma_1; \mu_2, \sigma_2; p$ 的混合高斯分布。其中，η 的分布列如下：

表 1: η 分布列

η	0	1
P	$1-p$	p

问题一

- 自己设定参数，用计算机生成10000 个混合高斯分布的随机数；
- 画出其频率分布直方图；
- 讨论不同参数对其分布“峰”的影响。

问题二

自己设定参数，用计算机生成1000 组，每组n 个混合高斯分布的随机数，设第i 组的随机数为$Z_{i,1}, Z_{i,2}, ..., Z_{i,n}$，且给出

$$U_i = \frac{\sum_{j=1}^{n} Z_{i,j} - n \cdot E(Z)}{\sqrt{n \cdot D(Z)}} \tag{1}$$

- 画出$U_1, U_2, ..., U_{1000}$ 的频率分布直方图；
- 讨论$n = 10, 20, 50, 100, 1000$ 对频率分布直方图“峰”的影响。
- 你能从中得出何结论。

2 关于混合高斯分布的简单讨论

我们首先讨论混合高斯分布的分布函数以及概率密度函数，来大致了解一下混合高斯分布的一些基本性质以及规律。

$$F_Z(z) = P(X + \eta Y \leq z) = P(X \leq z | \eta = 0) \cdot P(\eta = 0) + P(X + Y \leq z | \eta = 1) \cdot P(\eta = 1)$$

由于X,Y,η 相互独立，上式可化为：

$$F_Z(z)=(1-p)P(X\le z)+pP(X+Y\le z)=(1-p)F_X(z)+pF_{X+Y}(z) \tag{2}$$

其中，$F_{X+Y}(\cdot)$ 为$X+Y$ 的分布函数，由结论知，$X+Y\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$。于是我们得到了$Z$ 的分布函数如(3) 式所示，再在(2) 中两边对z 求导即得Z 的概率密度函数。

$$f_Z(z)=\frac{\mathrm{d}}{\mathrm{d}z}F_Z(z)=(1-p)\frac{\mathrm{d}}{\mathrm{d}z}F_X(z)+p\frac{\mathrm{d}}{\mathrm{d}z}F_{X+Y}(z)=(1-p)f_X(z)+pf_{X+Y}(z) \tag{3}$$

在(3) 式中代入$f_X(\cdot),f_{X+Y}(\cdot)$ 的表达式，有

$$f_Z(z)=(1-p)\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(z-\mu_1)^2}{2\sigma_1^2}}+p\frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}e^{-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}} \tag{4}$$

3 问题一的解答

3.1 服从混合高斯分布的随机数的生成

算法 1 (舍选法[1]) 设$f(\cdot)$ 为密度函数，且$\sup f(x)=f_0<\infty$，且满足$f(x)=0(\forall x\notin[a,b])$，那么生成以$f(\cdot)$ 为密度函数的随机变量X 的舍选算法如下：

1. 生成$[0,1]$ 上的独立均匀随机数U 和V，令$u=a+(b-a)U,v=f_0V$；
2. 若$v\le f(u)$ 则输出u；否则转1。

算法1 给出了一种给出密度函数$f(\cdot)$ 生成随机数的方法。容易将其应用在混合高斯分布的生成上，由于混合高斯分布中，$\lim_{x\to\infty}f(x)=0$，所以可以取一个很小的a和很大的b（如$a=-10^9,b=10^9$），则可以近似满足舍选法的条件，从而利用舍选法产生随机数。

舍选法的相关正确性的证明可以参见参考文献[1]，这里不再赘述。

若对于区间$[a,b]$ 中的大部分x，$f(x)<\varepsilon$，其中ε 为一个小量，则舍选法的步骤2 大概率会转到步骤1 继续选取，从而效率较为低下。针对于混合高斯分布，不仅算法效率较低，而且由式(4)，密度函数的描述过于繁琐，因此不作为生成服从混合高斯分布的随机数的首选。

在实际实现过程中，如果我们可以生成服从正态分布的随机数以及服从均匀分布的随机数，我们就可以利用下面这个简单的方法来获得产生随机数。

算法 2 生成服从参数为$\mu_1,\sigma_1;\mu_2,\sigma_2;p$ 的混合高斯分布的随机数的算法如下：

1. 生成满足$X\sim N(\mu_1,\sigma_1)$ 的随机变量X；生成满足$Y\sim N(\mu_2,\sigma_2)$ 的随机变量Y；
2. 生成满足$\eta'\sim U(0,1)$ 的随机变量η'，如果$\eta'\le p$，则令$\eta=1$；否则$\eta=0$；
3. 令$Z=X+\eta Y$，则Z 为服从参数为$\mu_1,\sigma_1;\mu_2,\sigma_2;p$ 的混合高斯分布的随机数。
4. 如果生成的随机数数量满足要求则退出，否则转1。

下面我们简要说明一下这个算法的正确性：首先，由于在混合高斯分布中，X,Y,η 三个随机变量相互独立，因此可以独立生成；其次，按照算法流程，生成的随机变量X,Y 可以满足$X\sim N(\mu_1,\sigma_1),Y\sim N(\mu_2,\sigma_2)$；然后，我们先生成了一个$(0,1)$ 内均匀分布的随机变量η'，那么，由随机分布的分布函数知，$P(\eta'\le p)=F_{\eta'}(p)=p,P(\eta'>p)=1-P(\eta'\le p)=1-p$，因

此按照算法2 所述规则生成的随机变量η 满足表1 所示的分布列，从而生成的随机数$Z = X + \eta Y$ 服从参数为$\mu_1, \sigma_1; \mu_2, \sigma_2; p$ 的高斯混合分布。

由于Matlab 中内置了生成正态分布和平均分布的函数，因此我们可以利用Matlab 非常方便的实现算法2，代码“GM_gen.m”如下：

```
close all; clear all; clc

% generate X from normal distribution
mu1 = input('Please input mu1: ');
sigma1 = input('Please input sigma1: ');
X = normrnd(mu1, sigma1, [10000 1]);

% generate Y from normal distribution
mu2 = input('Please input mu2: ');
sigma2 = input('Please input sigma2: ');
Y = normrnd(mu2, sigma2, [10000 1]);

% generate eta
p = input('Please input p: ');
eta_prime = unifrnd(0, 1, [10000 1]);
eta = zeros(10000, 1);
eta(eta_prime ≤ p) = 1;

% generate Z = X + eta * Y
Z = X + eta .* Y;

% graphing
[counts, centers] = hist(Z, 70);
figure
bar(centers, counts / sum(counts))

% printing data to file
fp = fopen('gaussian_mixture_data.csv', 'w');
for i = 1 : 10000
    fprintf(fp, '%d,\n', Z(i, 1));
end
fclose(fp);
```

设置参数为$\mu_1 = 0, \sigma_1 = 2; \mu_2 = 10, \sigma_2 = 5; p = 0.6$，则上述代码运行后产生的10000 组随机数见data 文件夹下的gaussian_mixture_data.csv 文件，所绘制的频率分布直方图如图1 所示。

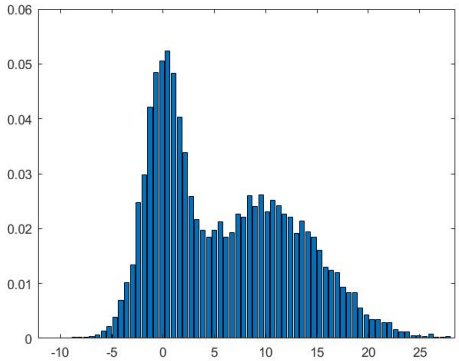
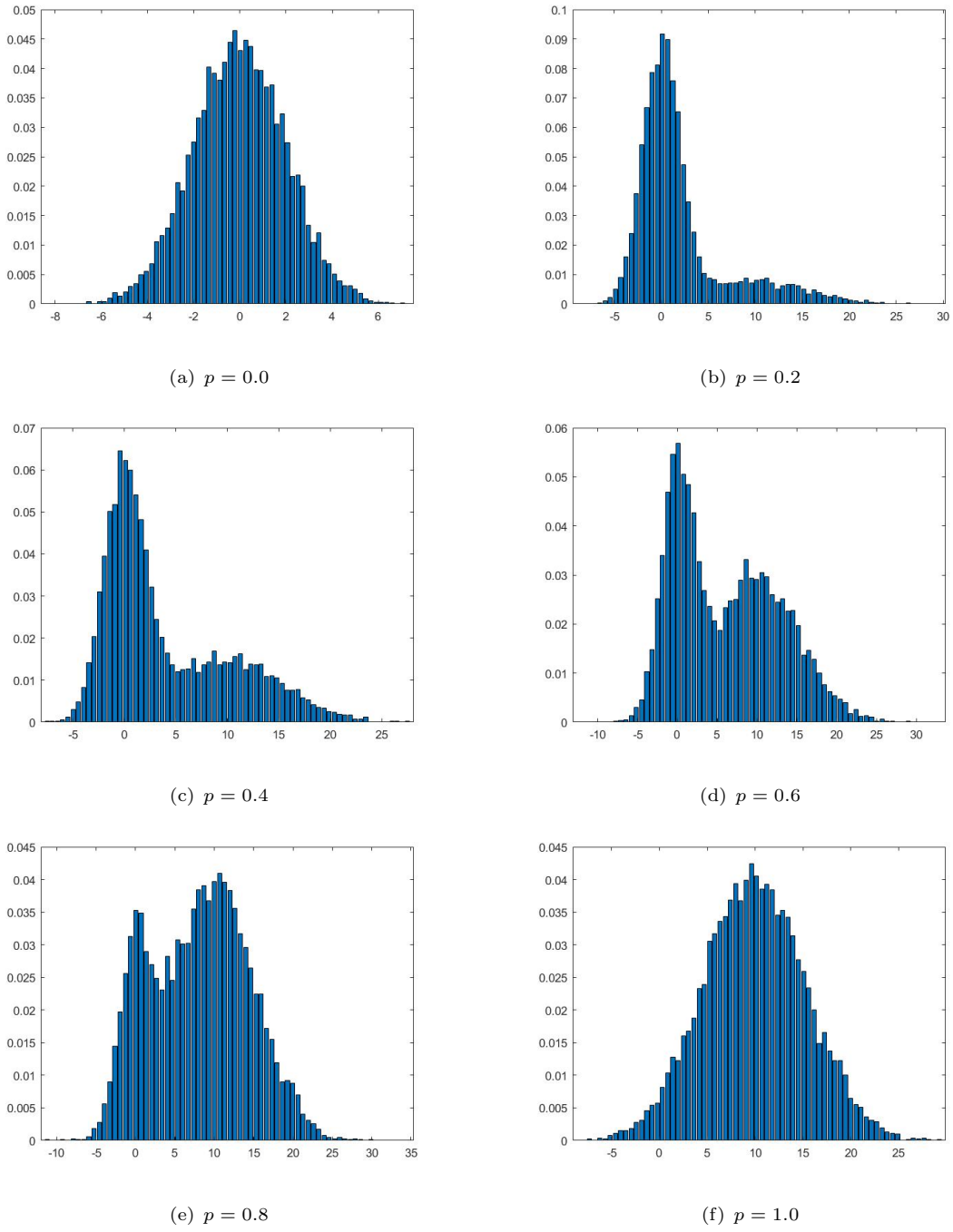


图 1: 参数为$\mu_1 = 0, \sigma_1 = 2; \mu_2 = 10, \sigma_2 = 5; p = 0.6$ 的高斯混合分布的随机数频率分布直方图

3.2 讨论参数对于混合高斯分布的影响

从式(4)中的密度函数我们可以看出，混合高斯分布的密度函数实际上是两个正态分布密度函数的加权平均，其中权重分别为p 和$(1-p)$；因此权重p 对于混合高斯分布的影响非常重要，因此我们先来讨论参数p 对于混合高斯分布的影响。

我们选取$\mu_1 = 0, \sigma_1 = 2; \mu_2 = 10, \sigma_2 = 5$，分别用上述Matlab 代码在$p = 0, p = 0.2, p = 0.4, p = 0.6, p = 0.8$ 和$p = 1.0$ 时各生成了10000 组随机数，可参见data 文件夹下的data2-0.csv, data2-1.csv, ..., data2-5.csv 文件，并分别绘制了他们的频率分布直方图如下：



(a) $p = 0.0$ (b) $p = 0.2$

(c) $p = 0.4$ (d) $p = 0.6$

(e) $p = 0.8$ (f) $p = 1.0$

图 2: 不同参数p 下生成的随机数的频率分布直方图

由于样本足够大，可以由频率估计概率，因此频率分布直方图的边缘曲线和混合高斯分布的

概率密度曲线近似重合，故从图2 中我们也可以观察不同参数下混合高斯分布的概率密度曲线。

当$p = 0.0$时，显然$Z = X \sim N(\mu_1, \sigma_1^2)$，图像只有单峰且峰值在$\mu_1$附近，为正态分布；当$p = 1.0$时，显然$Z = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$，图像也只有单峰且峰值在$\mu_1 + \mu_2$附近，为正态分布；观察图像可知，理论推导和实际结果相吻合。

当$0 < p < 1$时，图像出现了两个不同的峰值，分别在μ_1 和$\mu_1 + \mu_2$ 附近，且随着p 的增大，μ_1 这个峰的高度逐渐降低，$\mu_1 + \mu_2$ 这个峰的高度逐渐增高；即μ_1 附近的点的概率密度下降，而$\mu_1 + \mu_2$ 附近的点的概率密度上升。事实上，由式(4) 我们也可以清晰的认识到，其概率密度函数由两个正态分布函数以权重p和$1 - p$加权平均后构成，因此当p增大时，加权p的一项正态分布的参数为$\mu_1 + \mu_2$，这个正态分布出现的概率增大，因此$\mu_1 + \mu_2$ 这个峰的高度增高；同时，加权为$1 - p$的一项正态分布参数为μ_1，这个正态分布出现的概率减小，因此μ_1 这个峰的高度降低。这个理论推导的结果同样与图2 相符。

结论 1 一般地，当σ_1, σ_2较小时，参数μ_1, μ_2, p 决定了峰的位置、数量，当$p = 0$ 或$p = 1$ 或$\mu_2 = 0$ 时，图像仅有一个峰在μ_1 附近；否则，图像有两个不同的峰，分别在μ_1 附近和$\mu_1 + \mu_2$ 附近。

结论 2 若图像有两个峰，参数p 决定了两个峰的高度变化关系，当p 增大时，μ_1 峰的高度降低，$\mu_1 + \mu_2$ 峰的高度增高。

由于图像的峰对应着概率密度函数的极大值，因此我们可以将结论1, 2 表述如下：

结论 3 一般地，当σ_1, σ_2较小时，参数μ_1, μ_2, p 决定了混合高斯分布中概率密度函数的极大值点的位置、数量，当$p = 0$ 或$p = 1$ 或$\mu_2 = 0$ 时，概率密度函数仅有一个极大值点μ_1；否则，概率密度函数有两个不同的极大值点，分别为μ_1 和$\mu_1 + \mu_2$。

结论 4 若概率密度函数有两个极大值，参数p 决定了两个极大值的变化关系，当p 增大时，$f(\mu_1)$ 减小，$f(\mu_1 + \mu_2)$ 增大。

由于混合高斯分布本质上为两个正态分布$N(\mu_1, \sigma_1^2), N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ 的加权平均，因此参数σ_1, σ_2 的作用与正态分布基本相同，有：

结论 5 参数σ_1, σ_2 表征峰的陡峭程度，σ_1 越小，μ_1 附近的峰与$\mu_1 + \mu_2$ 附近的峰都越陡峭；σ_2 越小，$\mu_1 + \mu_2$ 附近的峰越陡峭，而不改变μ_1 附近的峰的陡峭程度。特别地，当参数σ_1, σ_2 过大时，两理论峰均过于平缓，因此在图像中无法明显体现出两个“峰”，而只能观察到一个“峰”如图3所示。

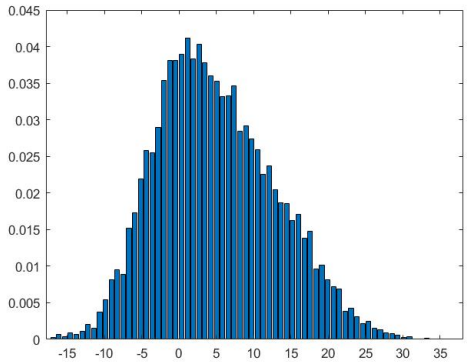


图 3: 参数为$\mu_1 = 0, \sigma_1 = 5; \mu_2 = 10, \sigma_2 = 5; p = 0.5$ 的高斯混合分布的随机数频率分布直方图

4 问题二的解答

在问题二中，我们选用参数$\mu_1 = 0, \sigma_1 = 2; \mu_2 = 10, \sigma_2 = 3; p = 0.6$。利用Matlab，按照第三节的方法可以生成服从混合高斯分布的矩阵，再对每行计算U_i，代码“GM_gen_U.m”如下：

```
close all; clear all; clc

% input n
n = input('Please input n:');

% initializing data
mu1 = 0; sigma1 = 2;
mu2 = 10; sigma2 = 3; p = 0.6;

% generating Z = X + eta * Y
X = normrnd(mu1, sigma1, [1000 n]);
Y = normrnd(mu2, sigma2, [1000 n]);
eta_prime = unifrnd(0, 1, [1000 n]);
eta = zeros(1000, n);
eta(eta_prime <= p) = 1;
Z = X + eta .* Y;

% computing E(Z_i) and D(Z_i) for each group
EZ = mean(mean(Z));
DZ = 0;
for i = 1 : 1000
    for j = 1 : n
        DZ = DZ + (Z(i, j) - EZ) .^ 2;
    end
end
DZ = DZ / (n * 1000);

% calculating U
for i = 1 : 1000
    tem = 0;
    for j = 1 : n
        tem = tem + Z(i, j);
    end
    U(i) = (tem - n * EZ) / sqrt(n * DZ);
end


% graphing
[counts, centers] = hist(U, 100);
figure
bar(centers, counts / sum(counts))

% printing data to file
fp = fopen('U_data.csv','w');
for i = 1 : 1000
    fprintf(fp, '%d,', U(i));
    for j = 1 : n
        fprintf(fp, '%d,', Z(i, j));
    end
    fprintf(fp, '\n');
end
fclose(fp);
```

首先我们解释一下为什么选用了如上参数。由问题一的讨论我们知道，在问题二中，选择的参数应该能够尽可能反映出混合高斯分布的特点，即Z 的频率分布直方图可以出现两个“峰”。因此我们选用了一系列参数，并用问题一的代码生成了10000 个随机数存储在data 文件夹下的data4.csv 文件中，并画出了Z 的频率分布直方图如图4 所示。

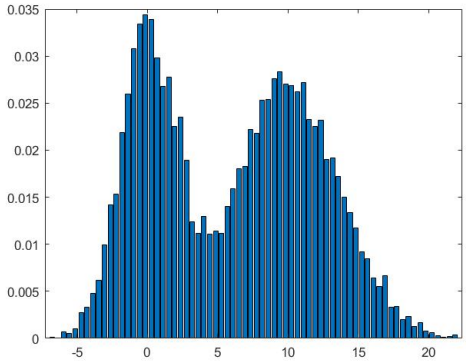
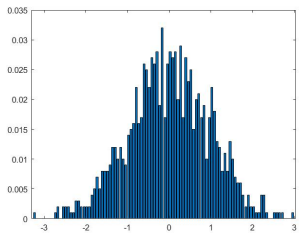
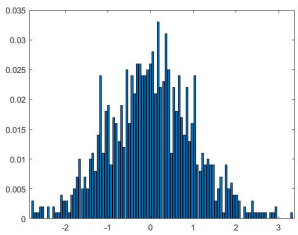


图 4: 参数为$\mu_1 = 0, \sigma_1 = 2; \mu_2 = 10, \sigma_2 = 3; p = 0.6$ 的高斯混合分布的随机数频率分布直方图

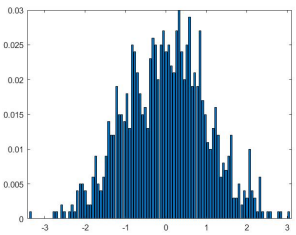
对于问题二，我们分别选取了$n = 10, n = 20, n = 50, n = 100$ 和$n = 1000$ 进行生成并计算，得到的数据全部存储在data 文件夹下的data5-1.csv, data5-2.csv, ..., data5-5.csv 中，并画出U_i 的频率分布直方图如图5 所示。



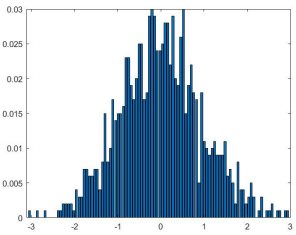
(a) $n = 10$



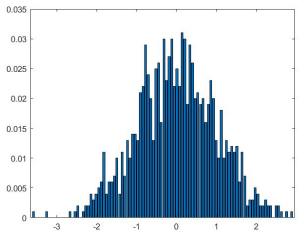
(b) $n = 20$



(c) $n = 50$



(d) $n = 100$



(e) $n = 1000$

图 5: 不同参数n 下U_i 的频率分布直方图

我们似乎并不能从中看出峰之间的明显的差距。

参考文献

[1] 李曙雄,杨振海. 舍选法的几何解释及其应用[J]. 数理统计与管理(4):40-43.